STIC-ILL

vol NO

From: Davis, Natalie
Sent: Tuesday, January 22, 2002 1:10 PM
To: STIC-ILL
Subject: 09/823101

Please send the following:

1. *Kato, et al., Nucleic Acids Res 1983 Dec 10;11(23):8197-203*

2. *Meyers, et al. J Biol Chem 1983 Aug 25;258(16):10128-35*

Natalie A.Davis, PhD
Patent Examiner
Art Unit 1642
CM1, Rm 8B13
Mailbox 8E12
Ph (703) 308-6410

379,615

L.C - NOS
1/23 - RC

WTS

AUGUST 25, 1983 VOLUME 258 NUMBER 16

ISSN 0021-9258
JBCHA3 258(16) 9581-10192 (1983)

THE Journal of Biological Chemistry

Published by The American Society of Biological Chemists, Inc.

FOUNDED BY CHRISTIAN A. HERTER
AND SUSTAINED IN PART BY THE CHRISTIAN A. HERTER MEMORIAL FUND

Vol. 258, No. 16

The Journal of Biological Chemistry

August 25, 1983

CONTENTS*

COMMUNICATIONS

* The CONTENTS arranged by Subject Categories will be found immediately following these CONTENTS.
Full *Instructions to Authors* will be found in THE JOURNAL, 258, 1 (1983), and reprints may be obtained from the editorial office.

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 258, No. 16. Issue of August 25, pp. 10128–10135. 1983
Printed in U.S.A.

Analysis of the 3′ End of the Human Pro-α2(I) Collagen Gene

UTILIZATION OF MULTIPLE POLYADENYLATION SITES IN CULTURED FIBROBLASTS*

(Received for publication, January 21, 1983)

Jeanne C. Myers‡, Leon A. Dickson§, Wouter J. de Wet‡¶, Michael P. Bernard‖, Mon-Li Chu‡, Maurizio Di Liberto§, Guglielmina Pepe‖, Frank O. Sangiorgi‡, and Francesco Ramirez‡‖

From the Departments of ‡Biochemistry and ‖Obstetrics and Gynecology, University of Medicine and Dentistry of New Jersey, Rutgers Medical School and the §Department of Biochemistry, University of Medicine and Dentistry of New Jersey, New Jersey School of Osteopathic Medicine, Piscataway, New Jersey 08854

Three overlapping genomic clones covering 28 kilobases of the human pro-α2(I) collagen gene have been isolated from a λ phage library. The analysis of 12 introns and 12 exons in the 3′ end region has shown that the human gene has a structure remarkably similar to that reported for the homologous chicken gene. One large intron, in the α-chain domain, contains an *AluI* sequence flanked by short direct repeats; a second *AluI* sequence is present 4 kilobases downstream from the termination codon. The analysis of the exon coding for the 3′-untranslated region has revealed that the pro-α2(I) collagen gene transcribes at least four different mRNAs in cultured fibroblasts. The colinearity and exact location of the termini of these transcripts was determined by Northern blots, R-looping analysis, S1 protection, and DNA sequencing. The ends of two transcripts are closely preceded by the canonical polyadenylation signal (AAUAAA), whereas two of its variations (AUUAAA and AUUAA) precede the ends of the other two transcripts.

The structural integrity of most organs and tissues depends on the harmonious expression of a complex battery of genes, including the multigene family encoding the different collagens. The developmentally regulated expression of the collagen genes results in the synthesis of at least nine different products which are subjected to a complex array of post-translational modifications and extracellular processing to produce the five different types of mature collagens known in vertebrates. The native proteins (procollagens) consist of three identical or similar pro-α-chains each with an NH_2-terminal propeptide, a COOH-terminal propeptide, and a central triple helical α-chain domain with a repetitive tripeptide structure $(Gly, X, Y)_{333}$. In the fibrillar collagens (types I–III), specific amino- and carboxyendopeptidases cleave extracellularly the propeptide segments before the mature proteins undergo the process of fiber formation (1). Alterations in the structure, synthesis, or processing of these proteins in man may result in a number of inherited disorders, such as osteogenesis imperfecta, chondrodystrophy, Marfan syndrome, and Ehlers-Danlos syndrome (2). One of our primary goals was to isolate and analyze, in detail, the structure of the human type I procollagen genes, in order to begin to understand those factors involved in their complex coordinated expression in normal and diseased tissues.

Type I procollagen, the most abundant of the five different collagens identified in higher vertebrates, is a major component of skin, tendons, and bones. This heterotrimer consists of two identical pro-α1(I) chains and one pro-α2(I) chain, and is, therefore, the product of two coordinately expressed genes (1). These two genes have been recently assigned to chromosome 7 (pro-α2(I)) and chromosome 17 (pro-α1(I)) (3, 4) using cloned cDNAs specific for these two chains (5, 6). Sequencing of the pro-α2(I) cDNA clones has allowed, for the first time, the determination of the primary structure of more than half of the human pro-α2(I) chain (7). The comparison of the human sequences with the previously published data on the homologous avian chain (8, 9) has made possible the examination of the evolution of this gene in two species which have diverged more than 300 million years ago (10).

Here, we report the isolation of three overlapping genomic clones covering 28 kb[1] of the human pro-α2(I) collagen gene, from its 3′ end to amino acid 19 in the helical portion of the α2(I) chain. The human pro-α2(I) gene exhibits a complexity of intron-exon organization analogous to collagen genes of other vertebrates, particularly in the size and distribution of the four exons coding for the COOH-propeptide (9, 11–14).[2,3]

Two repeated sequences, members of the *AluI* family, are associated with the human pro-α2(I) collagen gene. The first *AluI* sequence (α2R1) is present in a large intron between amino acid residues 765–766 in the α-chain. The second *AluI* sequence (α2R2) is located in the 3′-flanking region, 4 kb downstream from the termination codon. Both repeats are preceded by a 5′ poly(T) stretch and are flanked by a number of short direct repeats.

The detailed characterization of the first exon of the human pro-α2(I) gene has revealed the presence of multiple transcripts in cultured fibroblasts. We were able to characterize at least four different transcripts, varying in the length of their 3′-untranslated regions. Two of these utilize the canonical polyadenylation signal (AAUAAA), whereas the other two utilize two variations of it (AUUAAA and AUUAA). Multiple transcripts with similar characteristics have already been described for other eukaryotic genes (15–19) and in one case they have been correlated to tissue specificity (20).

* This work was supported by Grants AM 16, 516-C05, AM 30387-01, and RR 09085 from the National Institutes of Health and the Lalor Foundation. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

¶ Sponsored by a grant from the South African Medical Research Council. Present address, Department of Biochemistry, Potchefstroom University, Potchefstroom 2520, South Africa.

[1] The abbreviations used are: kb, kilobases; bp, base pairs.

[2] M. P. Bernard, M. L. Chu, J. C. Myers, F. Ramirez, E. Eikenberry, and D. J. Prockop, manuscript in preparation.

[3] Y. Yamada, personal communication.

MATERIALS AND METHODS

Enzymes and Isotopes—Restriction endonucleases and other nucleic acid-modifying enzymes were purchased from Bethesda Research Laboratories and New England Biolabs (Beverly, MA) and used according to the manufacturer's specifications. S1 nuclease was purchased from Miles Laboratories Inc. (Elkhart, IN), ultrapure formamide from Fluka. Labeled isotopes were purchased from New England Nuclear and Amersham Corp., nitrocellulose filters from Schleicher & Schuell.

Screening of Genomic Library—The genomic library in Charon 4A used in these studies was obtained from Dr. T. Maniatis (Harvard University) and contained 15-20-kb inserts of human nuclear DNA partially digested with the enzymes *Alu*I and *Hae*III (21). The screening for the pro-α2(I) collagen clones, their isolation, amplification, and DNA purification were performed as described (22). All the experiments were conducted using the appropriate level of biological and physical containment as detailed in the National Institutes of Health guidelines for recombinant DNA research.

RNA Isolation and DNA Sequencing—Total poly(A$^+$) RNA was purified from cultured human fibroblasts as previously described (23). The nucleotide sequences of the appropriate DNA fragments were carried out according to the chemical modification procedure of Maxam and Gilbert (24). Sequencing of both strands was performed for most of the regions detailed in this paper.

Hybridization to RNA and DNA Immobilized onto Nitrocellulose—Total poly(A$^+$) RNA was electrophoresed in 0.7% agarose gels in 2 M formaldehyde and transferred at 4 °C by blotting onto nitrocellulose paper (25). Restricted DNA in agarose gels was denatured and neutralized *in situ* and transferred onto nitrocellulose paper using the techniques of Southern (26). Nucleic acid filter-bound hybridizations were performed as previously described (5, 6).

Electron Microscopy—R-looping was carried out according to the protocol of Kaback *et al.* (27), and heteroduplexing according to the method described by Davis *et al.* (28). DNA molecules were visualized and photographed with a JEOL electron microscope and measured at a final magnification of 45,000 with a Hewlett-Packard 9810 calculator equipped with a 9864A digitizer. Double-stranded replicative form of phage φX174 DNA was included for length calibration.

S1 Nuclease Protection—End-labeled genomic fragments were used for the S1 protection experiments using the protocol described by Berk and Sharp (29). The 3' end labeling was carried out by the addition of [α-^{32}P]deoxynucleotides (specific activity: 1,000 Ci/mM) using the Klenow fragment of DNA polymerase I. The labeled fragments were heat-denatured in 30% dimethyl sulfoxide, strand-separated on polyacrylamide gels, electroeluted, and annealed to total fibroblast poly(A$^+$) RNA prior to S1 digestion. The exact sizes of the S1-resistant products were determined by electrophoresis on a 5% sequencing gel (80 cm) in parallel with DNA fragments which were 5' end-labeled and subjected to the Maxam and Gilbert (24) chemical modification reactions.

RESULTS AND DISCUSSION

Gene Isolation—The pro-α2(I) cDNA clones, Hf-32, and Hf-1131 (5, 7) were used for the initial screening of the genomic library. A positive clone (NJ-1), 16.8 kb in length, was isolated and appropriate subclones were subsequently used for the isolation of 5' end (NJ-3) and 3' end (NJ-6) overlapping genomic clones. The three clones (40 kb in total length) were extensively characterized by restriction endonuclease mapping and Southern blot hybridization with different subfragments of Hf-32 and Hf-1131 in order to define their sequential orientation. The continuity of the overlapping genomic regions was confirmed by Southern blot analysis of nuclear DNA digested with various restriction enzymes. The presence of repeated sequences associated with the pro-α2(I) gene was determined by Southern blot hybridization of the three clones with "nick-translated" total human DNA. Electron microscopy studies and DNA sequencing were performed for the portions of the gene which are the subject of the investigations presented here.

A composite restriction map of the human pro-α2(I) collagen gene with its relationship to the different domains of the protein is depicted in Fig. 1. The clones span from amino acid 19 in the α-chain to the 3'-flanking region. They cover 28 kb of the pro-α2(I) collagen gene and contain almost 4 kb of coding sequences. This ratio of interdispersion with noncoding sequences is almost identical with that found by Wozney *et al.* (13) for the pro-α2(I) chicken gene. We, therefore, can safely extrapolate that the size of the entire human pro-α2(I) collagen gene should not significantly vary from the 38-kb value reported for the avian gene.

Analysis of the 3' End of the Gene: Exon-Intron Arrangement—The complexity of the chicken pro-α2(I) collagen gene has been documented by a series of elegant investigations which have shown that this coding unit is greatly interdispersed by almost 50, often very large, introns resulting in a gene exceeding at least eight times the size of its mature transcript (for a review see Tate *et al.* (30)). The function of introns in eukaryotic genes is still a subject of speculation; one of the theories favors the idea that they separate exons encoding different functional or conformational segments within the same protein (31). This theory has an evolutionary significance, because it implies that complex proteins can evolve different functional domains independently of each other. Moreover, a particular function could be developed only once during evolution and then, through recombination and rearrangement of blocks of DNA, be dispersed among different genes. In line with this idea, Wozney *et al.* (13) have suggested that the four exons encoding the chicken pro-α2(I) collagen COOH-propeptide represent four different functional domains of this portion of the protein. The same group has also observed the absence of introns in the junction regions between the terminal propeptides and the helical portion of the α-chain. They have concluded that the junction exons represent evolutionary stable domains of the gene, because they encode for the endopeptidase cleavage sites of the fibrillar collagens.

The isolation of the human pro-α2(I) gene has now allowed us to compare these features in the mammalian gene. A map of the intron-exon arrangement in the 9 kb of the gene extending from residue 765 in the α-chain to the end of the 3'-untranslated region was determined by electron microscopy (Fig. 1). The approximate sizes of the introns and exons, as determined by electron microscopy, are summarized in Table I. A more detailed analysis of some sections of importance was obtained by DNA sequencing (Figs. 2 and 6). Twelve exons and 12 introns are present in this region and show a size and distribution remarkably similar to that reported for the chicken pro-α2(I) gene (30).

The first four exons (759 bp) encode primarily for the COOH-terminal propeptide and are interrupted by 2 kb of noncoding sequences distributed between three introns. Exon 1 codes for the last 48 amino acids of the COOH-terminal propeptide and contains the entire 3'-untranslated region. The complete sequence of this exon was determined (Fig. 6) and will be discussed in greater detail in a later section. Exon 2 contains the carbohydrate attachment site in a region which is highly conserved in the human[2] and chicken pro-α1(I) and pro-α2(I) genes (7, 8) as well as the avian pro-α1(III)[3] gene. Exon 3 contains the tricysteine cluster. Exon 4, the junction exon, codes for the end of the triple helical domain, the telopeptide, and the beginning of the COOH-propeptide. The remainder of the 12 exons shown in Fig. 1 code for the COOH-terminal 264 amino acid residues of the α-chain domain. These eight exons are small and are interrupted by seven introns, ranging between 100 and 1000 bp in size (Table I). A distinct pattern in the distribution of small and large exons in this region is evident from the data presented in Table I. This pattern closely resembles the arrangement of 54- and 108-bp exons in the same region of the chicken pro-α2(I) gene more accurately determined by direct DNA sequencing (30).

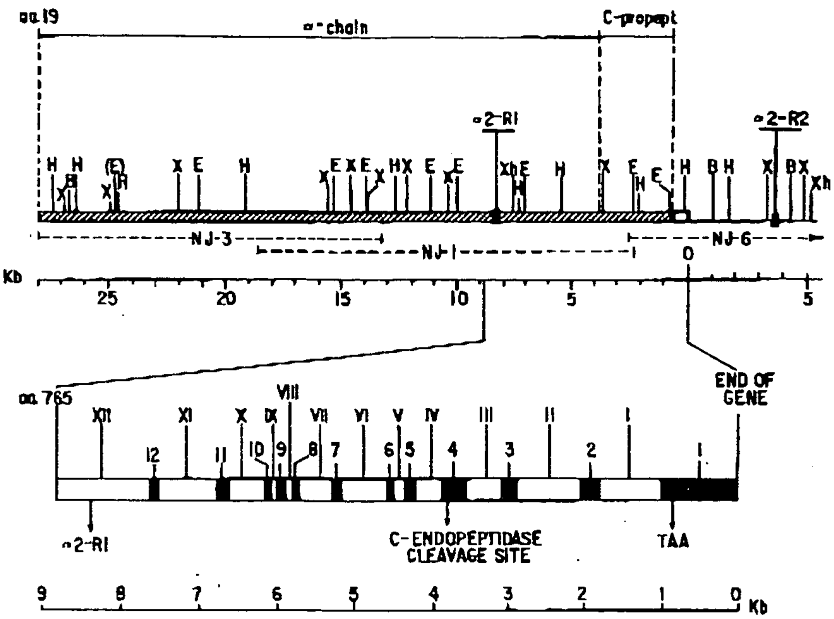

FIG. 1. **Restriction map of part of the human pro-α2(I) collagen gene, its relationship to the different domains of the polypeptide chain, and the intron-exon arrangement of the 3′ region of the gene.** The *upper half* shows 28 kb of gene (*cross-hatched box*) spanning from amino acid residue 19 in the α-chain to its 3′ end (position 0 in the scale *below*). The *white box* indicates the 3′-untranslated region. Depicted are also 5 kb of the 3′-flanking sequences which include the *AluI* repeat, *α2R2*. The other repeat, *α2R1*, is located within the gene 8.5 kb from its 3′ end. The *letters* represent all sites of the following restriction enzymes within the analyzed 28 kb of genomic DNA: *B*, *BamHI*; *E*, *EcoRI*; *H*, *HindIII*; *X*, *XbaI*; *Xh*, *XhoI*. The *EcoRI* site, located 25 kb from the 3′ end of the gene and indicated in *parentheses*, was found to be a polymorphic site in several individuals. The overlaps of the three genomic clones NJ-1, NJ-3, and NJ-6 are depicted *underneath* the restriction map (*broken lines*). NJ-6 contains an additional 7 kb of flanking sequences not shown in the figure. The *lower half* shows an expanded representation of the intron-exon arrangement of the gene from amino acid residue 765 in the α-chain to the 3′ end of the gene. The exons (*black boxes*) and the introns (*white boxes*) are sequentially numbered from the 3′ end with *Arabic* and *Roman numerals*, respectively. The locations of the termination codon (TAA), the carboxy endopeptidase cleavage site, and the α2R1 repeat are indicated in exons 1 and 4 and intron XII.

TABLE I

Exon-intron arrangement of the 3′ region of the human pro-α2(I) collagen gene

Exon/Intron	Size in bp
Exon 1	992[a,b]
Intron I	770[c]
Exon 2	243[c]
Intron II	900[c]
Exon 3	192 ± 30[d]
Intron III	482 ± 123[d]
Exon 4	286 ± 30[d]
Intron IV	393 ± 48[d]
Exon 5	118 ± 22[d]
Intron V	143 ± 23[d]
Exon 6	74 ± 9[d]
Intron VI	600 ± 85[d]
Exon 7	119 ± 16[d]
Intron VII	471 ± 80[d]
Exon 8	69 ± 10[d]
Intron VIII	132 ± 26[d]
Exon 9	113 ± 11[d]
Intron IX	98 ± 22[d]
Exon 10	72 ± 8[d]
Intron X	467 ± 19[d]
Exon 11	132 ± 19[d]
Intron XI	747[c]
Exon 12	113[c]
Intron XII	1052[a]

[a] The sizes were determined by direct DNA sequencing.
[b] Exon 1 contains 144 bp of coding sequences (Fig. 6).
[c] The sizes were determined by electron microscopy on one molecule.
[d] The sizes were determined by electron microscopy on at least 15 molecules, and the standard deviations are indicated.

In this context, it must be noted that the data in Table I reflects the real sizes of the exons only approximately. Our preliminary sequence data showed that exons 6, 8, and 10 are indeed 54 bp and that the size of exons 5, 7, 9, 11, and 12 is 108 bp.[4] Although numerous differences have been found at the nucleotide level between the human and chicken pro-α2(I) collagen cDNAs (7-9), our data indicate that the structure of the 3′ end of the gene is almost identical. The similarity is evident both at the level of the overall intron-exon arrangement and in the distribution of the four coding segments within the COOH-propeptide region. It is interesting to note that both the chicken pro-α1(III)[3] and the human pro-α1(I)[5] collagen genes have the same number of exons coding for the COOH-propeptide. Therefore, at least for the COOH-propeptide region, these observations seem to favor the functional domain hypothesis. However, this hypothesis does not explain the separation of the α-chain region into 40 exons, even by postulating the presence of clusters of functional subdomains (32). The detailed analysis of this particular section of the collagen gene and the molecular characterization of the defect in those patients, where structural abnormalities of the α-chain are due to either insertions or deletions (33–35), may in the future help answer some of these questions.

The analysis of the 40 kb of genomic DNA covered by the three overlapping clones has also revealed the presence of two short repeated sequences, which have been mapped by South-

[4] M. Di Liberto, V. Benson, R. Shemesh, T. Mariano, and L. A. Dickson, manuscript in preparation.
[5] M. L. Chu, W. de Wet, M. P. Bernard, M. Morabito, J. C. Myers, C. J. Williams, and F. Ramirez, manuscript in preparation.

FIG. 2. **The nucleotide sequence of the *AluI* repeats associated with the human pro-α2(I) collagen gene.** *Top,* the nucleotide sequence of the *AluI* repeated sequence, α2R1, located in intron XII. Flanking α2R1 are four direct repeats which are *underlined* and identified with *Roman numerals. Bottom,* the nucleotide sequence of the *AluI* repeated sequence, α2R2, located 4 kb downstream from the termination codon. Flanking α2R2 are five direct repeats which are *underlined* and identified with *Roman numerals.*

α2R1

```
 I        II        III
GATATACTGATTTTACTCATTTCCTTTTTGTTTTTGTTTTTGTTTTGTTTTTGTTTTTTTCTTTTGGGATCGAGTTCCCCTCTGTGCC
CATGCTGCACTGCAGTGGCACGATCTCAGCTCACTGCGACTTCCCCCTCCCCAGTTCAAGTAATTATCTTGTCTCAGCCTCCCGACTACC
TGGGACTACAGGCACACACCATCATGCCTAGCTAATTTTCGTATTTTCAGTAAAGACCCCCTTTCACCATATTGGTCAGGCTGGTCTCGA
                                                                                      IV
TCTCCTGACCTCAGGTGATCCACCACCTTGACCTCCTAAAGTGCTGGGATTACACATGTGAGCCACCCCACCCAGCCTGATTTCCT
```

α2R2

```
 I        II       III
GACTTTCTGAACTATGACTTTCTCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTCATCACTTATACTCTTCTTCCACAGGCTGGAGTGCA
GTGGTGCGATCTTGGATCACTACAGCTTGAACTCCCAGGTTCAACCTACTCTACTACCTCATCCTCCCTAGTAGCTGCGATTACACCCAT
GCACCACCATGCCCAGCTAATCTGATATTTTTAGTAGAGACAGGTTGTGCCACGTTGGTCAGGCTGGTCTCGAACTCCTGACCTCAGGTG
                                                                    IV      V
ATCCGCCCTTCTCGGCCTCCCAAAGTGCTCGCATTACACCTCTCACCCACTGCACCCGGCTTGAACTATGACTTTCA
```

ern blotting hybridization with respect to the pro-α2(I) collagen gene (Fig. 1). Sequencing experiments have determined that they are two different members of the *AluI* family, which is found randomly interdispersed throughout the human genome (36). Several theories have been suggested to explain the function of these elements and the mechanisms behind their ubiquitous dispersal in the mammalian genome. One line of thought has correlated the presence of short direct repeats flanking the *AluI* to the direct repeats flanking bacterial transposable elements (37).

One of the pro-α2(I) *AluI* repeats (α2R1) is contained in intron XII between residues 765 and 766 in the α-chain, 315 nucleotides from the 5′ end of the intron and 430 nucleotides from the 3′ end (Fig. 1). The central *AluI* consensus region of α2R1, 281 bp long, is preceded at the 5′ end by a long stretch of poly(T), and it is flanked at both ends by short direct repeats (Fig. 2, *top*) (36). Three direct repeats (I–III) are present at the 5′ end and one (VI) is present at the 3′ end. Only direct repeats III and IV are identical. Direct repeat II differs from III by the insertion of a Thd and has a Cyd to Ado change, while direct repeat I differs from III by a Thd to Ado and a Cyd to Ado change.

The other pro-α2(I) *AluI* repeat (α2R2) is located in the flanking sequences 4 kb downstream from the termination codon. As for α2R1, α2R2 is an inverted *AluI* repeat with an 80% homology to the *AluI* consensus sequence and varies from α2R1 for both the internal segment and for the number of the flanking short direct repeats. Three direct repeats are present at the 5′ end (I–III) and two at the 3′ end (IV and V) (Fig. 2, *bottom*). These direct repeats are seven to nine nucleotides long and only two direct repeats, II and IV, are identical. Direct repeat III differs from I only by the presence of an extra Cyd at its 3′ end. In addition, the seven nucleotides of direct repeat II plus the first seven nucleotides of direct repeat III are identical with the corresponding first 14 nucleotides of direct repeats IV and V.

At the present time, one cannot draw any conclusions about the possible role that these *AluI* sequences might have played in the evolution of this gene; however, it is interesting to note that our studies show that two different *AluI* members are also closely associated with the human pro-α1(I) collagen gene,[5] which resides on a different chromosome (4). Unlike the *AluI* sequences described here, the *AluI* consensus sequence of the pro-α1(I) repeats are flanked at the 3′ end by a long poly(A) stretch (36). As for α2R2, one of the pro-α1(I) *AluI* repeats is located 4 kb downstream from the termination codon, whereas the other repeat is contained within an intron of the α-chain between amino acid residues 411–412. Finally, the two *AluI* repeats associated with each gene are spaced at an almost identical distance of 12 kb apart. One could speculate that the difference in the relative position of the intronic *AluI* repeats in the two genes may be due to the greater compactness of the human pro-α1(I) collagen gene which is only 18 kb in size.[5]

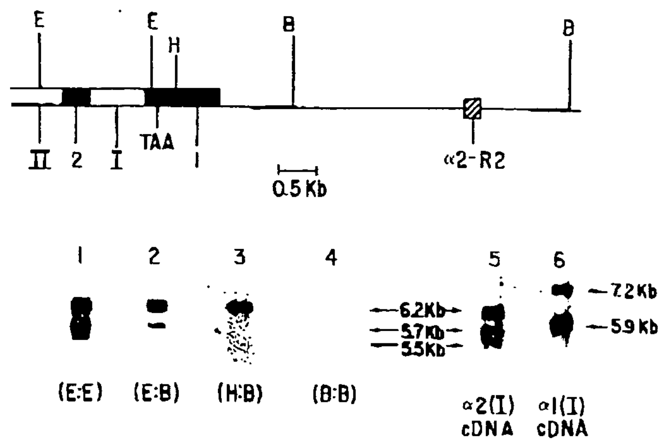

FIG. 3. **Northern blot hybridizations of total human poly(A⁺) RNA from cultured fibroblasts with various genomic probes.** *Upper half,* restriction map of a portion of the 3′ end of the pro-α2(I) collagen gene including part of the flanking sequences. Exons 1 and 2 are shown as *dark boxes*; introns I and II are shown as *white boxes*. The locations of the termination codon (*TAA*) and the α2R2 repeat (*cross-hatched box*) are provided as points of reference to the map shown in Fig. 1. The *letters* represent the sites of different restriction enzymes: *B, BamHI; E, EcoRI; H, HincII. Lower half,* Northern blot hybridizations of various genomic fragments with 0.5 μg of total poly(A⁺) RNA. *Lane 1,* the probe used was the 1.5-kb genomic fragment *EcoRI:EcoRI* (*E:E*). *Lane 2,* the probe used was the 1.8-kb genomic fragment *EcoRI:BamHI* (*E:B*). *Lane 3,* the probe used was the 1.5-kb genomic fragment *HincII:BamHI* (*H:B*). *Lane 4,* the probe used was the 3.6-kb genomic fragment *BamHI:BamHI* (*B:B*), which unlike other reports (46), under our experimental conditions, did not show any hybridization pattern to a mature pro-α2(I) collagen mRNA species even after long exposure. *Lanes 5* and *6,* hybridization with cDNA probes specific for the pro-α2(I) chain (Hf-32) and the pro-α1(I) chain (Hf-677). The size of the different RNA species was determined by running RNA markers in a parallel slot and visualizing them with ethidium bromide staining. The RNA markers used were: 35 S polio virus RNA (7.6 kb), 28 S (5.1 kb) and 18 S (2.15 kb) chicken fibroblast rRNA, and 23 S (3.3 kb) and 16 S (1.54 kb) *Escherichia coli* rRNA.

Polymorphic mRNAs—We have reported that when total poly(A⁺) RNA and polysomal RNA from normal human fibroblasts is blotted onto nitrocellulose paper and hybridized to the pro-α1(I) and pro-α2(I) cDNA clones, multiple bands of almost equal intensity are observed with each probe (Fig. 3) (5, 6). This finding indicates that both type I collagen genes transcribe more than one mRNA. We have found that this phenomenon is also true for mouse and chicken fibroblasts

(data not shown). The presence of polymorphic mRNAs due to either length heterogeneity or alternative splicing has been already described in several other eukaryotic genes (15–19). These transcriptional variations in the mouse α-amylase gene are related to tissue specificity (20) whereas, in the μ-immunoglobulin system and the calcitonin gene, they appear to cause the expression of functionally diverse products (38, 39). Quantitative and qualitative differences in the types and proportions of the different collagen proteins in various normal and abnormal tissues have been reported (1). In order to ultimately assign a biological and functional role to these mRNAs, we have addressed the basic question of the exact number and structural characteristics of the pro-α2(I) collagen gene transcripts in human fibroblasts.

First, we established the transition of the different RNAs at the 3′ end of the gene by Northern blot hybridizations with the appropriate genomic subclones. Second, we confirmed the colinearity of these transcripts by R-looping experiments between the genomic subclones and the collagen mRNAs. Third, we defined the exact location of the termini of the transcripts by S1 nuclease protection experiments in conjunction with DNA sequencing. The estimated size, of the three major mRNA bands observed by Northern blot hybridization with the pro-α2(I) cDNA probe are 6.2, 5.7, and 5.5 kb, respectively (Fig. 3). This pattern was consistently seen in a number of human fibroblast cell lines. However, the 5.5-kb band, which represents only 5–10% of the hybridizing RNA, was not detectable by Northern blot hybridization, R-looping analysis, or S1 nuclease protection experiments using genomic probes specific for the 3′-untranslated region. At this time, we do not have any conclusive explanation for the nature of this RNA species and for the location of its 3′ terminus. Currently, cDNA cloning experiments, aimed to isolate and directly characterize this particular transcript by DNA sequencing, are in progress.

The 5.7-kb mRNAs—The experiments summarized in Fig. 3 clearly show that the transition between the 5.7- and 6.2-kb mRNAs is located in the 3′-untranslated region of the pro-2(I) gene, more precisely in exon 1 around the *Hinc*II site. This observation was visually confirmed by R-looping experiments using the genomic *Eco*RI:*Bam*HI fragment subcloned in pBR322. Ten of the 41 R-loops analyzed showed a DNA:RNA hybrid of 384 ± 73 nucleotides (Fig. 4*A*). To determine the exact terminus of the 5.7-kb mRNA, S1 nuclease protection experiments were performed using as probe the *Eco*RI:*Ava*II (*E*:*A*) genomic fragment which extends 194 nucleotides beyond the *Hinc*II site (Fig. 5). This 487-nucleotide fragment was 3′ end-labeled, strand-separated, hybridized to total fibroblast poly(A⁺) RNA, and subjected to S1 digestion, and the product of the reaction was run on a polyacrylamide gel. A major S1-resistant product was seen as a close triplet of bands 345, 333, and 330 nucleotides long, designated in Fig. 5 as an average value of 340. This result placed the polyadenylation attachment site of the 5.7-kb mRNA within 20 nucleotides from the canonical AAUAAA signals (Fig. 6) and in accordance with the 384 ± 73 observed R-loop. The end of the 5.7-kb mRNA transcript is shown in Fig. 5 between position 307 and 285 from the termination codon. The 487-nucleotide-resistant product was the result of complete protection of the *Eco*RI:*Ava*II fragment by the 6.2-kb mRNA, which proved once more the colinearity of the major transcripts. Furthermore, after longer exposure, a minor S1-resistant species 250 nucleotides long was seen, accounting for less than 5% of the total protected material. The pentanucleotide AUUAA, a shorter variation of the canonical signal, closely precedes the end of this mRNA (Fig. 6). The fact that this minor transcript had not been seen by electron microscopy and Northern blots was probably due to its low representation and to the small difference in length with respect to the major 5.7-kb mRNA. We excluded that this minor S1-resistant species represents the 5.5-kb mRNA by performing Northern blot hybridizations with a 290-bp *Eco*RI:*Hinc*II genomic fragment (Figs. 5 and 6). This short genomic fragment covers 250 bp specific for the minor S1-

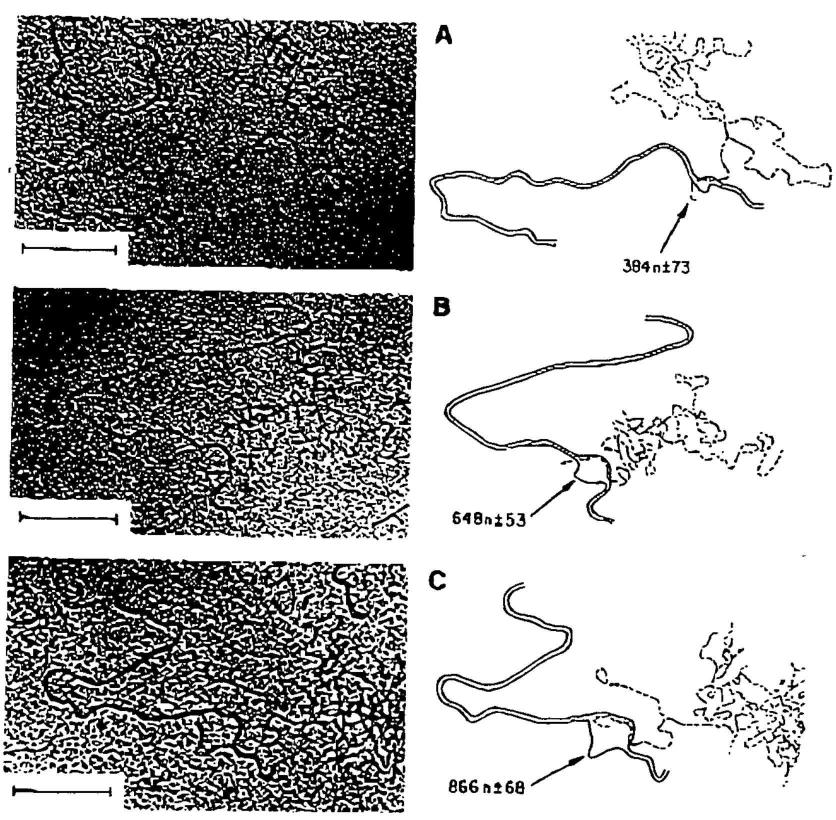

FIG. 4. **R-looping analysis of the 3′ end of the pro-α2(I) collagen gene.** The 1.8-kb genomic subclone *Eco*RI:*Bam*HI (which includes part of exon 1 and the 3′-flanking region; see maps in Figs. 1 and 3) was subcloned in pBR322, and linearized with the enzyme *Pvu*II prior to the R-looping hybridization. Three different types of DNA:RNA hybrid molecules were seen using the same genomic fragment, and they are shown in *A*, *B*, and *C*. The interpretive tracings of the three micrographs at the *left* are shown on the *right*, with the indication of the sizes of the R-loops. *Solid lines*, DNA; *broken lines*, RNA. The sizes of the three R-loops were determined using as standard the double-stranded replicative form of phage φX174 DNA. The *bar* in the *lower left corner* of the pictures is a length standard of 1.0 kb.

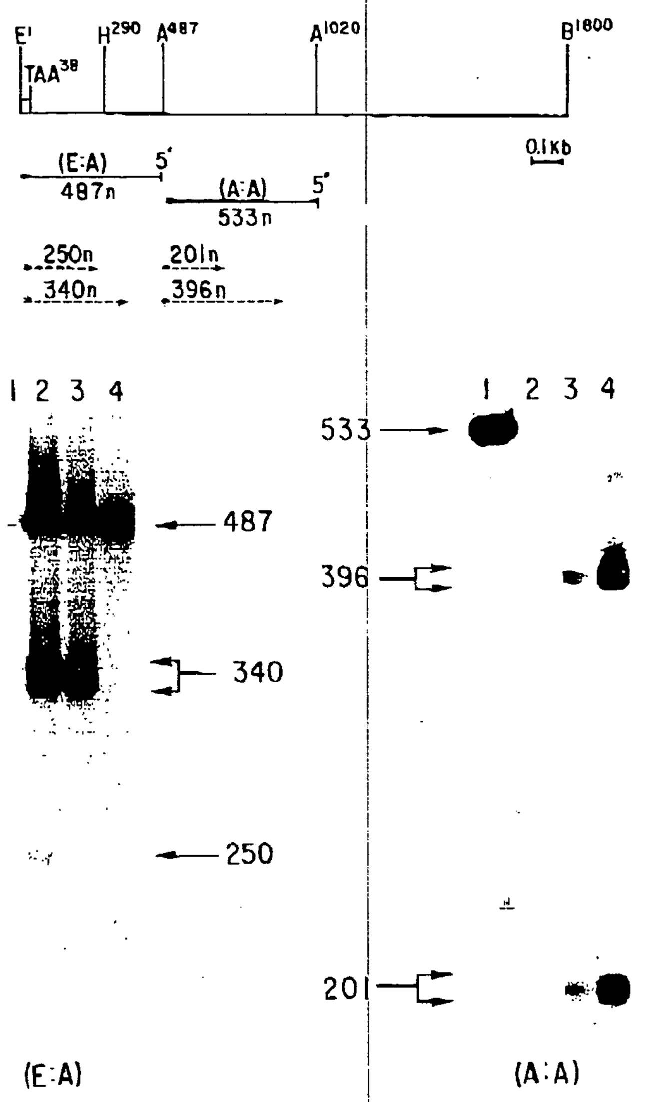

FIG. 5. **S1 nuclease mapping of the 3′ termini of human fibroblast pro-α2(I) collagen mRNAs.** *Upper half*, restriction map of the 3′ end of the gene and part of its flanking sequences. The *letters* designate the sites of the different restriction enzymes: *A*, *Ava*II; *B*, *Bam*HI; *E*, *Eco*RI; *H*, *Hinc*II. The *superscripts* indicate the nucleotide distance of the various cleavage sites from that of *Eco*RI (designated as 1). The termination codon (*TAA*) is shown at the *end* of the coding sequence (*white box*). The *solid lines* represent the DNA fragments used in these experiments; the *broken lines* represent the resistant products obtained after S1 digestion. The *asterisks* indicate the 3′ end labeling of the molecules. The two fragments used were *Eco*RI:*Ava*II (*E:A*), 487 nucleotides long, and *Ava*II:*Ava*II (*A:A*), 533 nucleotides long. Both fragments were 3′ end-labeled, strand-separated, and electroeluted. Constant amounts of the labeled antistrands (1 ng) were hybridized to increasing concentrations of total poly(A^+) RNA under R-loop conditions to minimize self-reannealing. The hybrids were then subjected to S1 nuclease digestion as described by Berk and Sharp (29). *Lower half*, the *left side* shows the autoradiogram of the S1 nuclease digestion using the 487-nucleotide *Eco*RI:*Ava*II (*E:A*) fragment. *Lane 1*, labeled DNA with no RNA and S1-treated; *Lane 2*, labeled DNA hybridized to 3 μg of total poly(A^+) RNA and treated with S1; *Lane 3*, labeled DNA hybridized to 1 μg of total poly(A^+) RNA and treated with S1; *lane 4*, labeled DNA with no RNA and no S1 treatment. The *right side* shows the autoradiogram of the S1 nuclease digestion using the 533-nucleotide *Ava*II:*Ava*II (*A:A*) fragment. *Lane 1*, labeled DNA with no RNA and no S1 treatment; *Lane 2*, labeled DNA with no RNA and treated with S1; *Lane 3*, labeled DNA hybridized to 1 μg of total poly(A^+) RNA and treated with S1; *Lane 4*, labeled DNA hybridized to 10 μg of total poly(A^+) RNA and treated with S1. The sizes of the S1-resistant products are indicated as averaged values. The *Ava*II:*Ava*II fragment contained a small, fast moving contaminant (*right*, *Lane 1*) which did not interfere with the S1 assay results.

resistant product and therefore is a more sensitive probe for hybridizing to the 5.5-kb RNA. In these experiments, the *Eco*RI:*Hinc*II probe detected only the 6.2- and 5.7-kb bands even under less stringent conditions of hybridization (data not shown).

The 6.2-kb mRNAs—The pattern obtained by Northern blot hybridization with the different 3′ end genomic fragments (Fig. 3) indicates that the end of the 6.2-kb band is located within the *Hinc*II:*Bam*HI region. Of the 41 DNA:RNA hybrids analyzed, 20 showed a hybrid 866 ± 68 nucleotides long and 11 showed a DNA:RNA molecule 648 ± 53 nucleotides long (Fig. 4, *B* and *C*). This finding suggested that the 6.2-kb band seen by Northern blot hybridization (Fig. 3) was indeed a composite of two mRNA species with different 3′ termini. The end of these two transcripts was determined by S1 nuclease protection experiments using the 533-nucleotide *Ava*II:*Ava*II fragment immediately adjacent to the 487 nucleotide *Eco*RI:*Ava*II fragment (Fig. 5). Two S1-resistant bands of almost equal intensity were seen: one, 201 nucleotides long, the other, a doublet averaging 396 nucleotides. The latter would place the end of this transcript around position 845 from the termination codon (Fig. 6), well in agreement with the presence in that area of the AAUAAA signals and in accordance with the 866 ± 68 R-loop. The former would place the end of the other transcript at position 650 from the termination codon or 20 nucleotides from an AUUAAA signal (Fig. 6), explaining the mRNA species seen as a 648 ± 53 R-loop.

From these data, we conclude that the 6.2-kb band seen by Northern blots is actually two co-migrating RNAs varying 200 nucleotides in the length of their 3′-untranslated region. One of the possible explanations is that the size difference at the 3′ end is compensated by a differential length of their 5′-untranslated regions. The isolation of the 5′ end of the pro-α2(I) collagen gene will allow us to test this hypothesis and to prove if these differences are due to either differential initiation or splicing.

Polyadenylation Sites—The role of the hexanucleotide AAUAAA in eukaryotic genes as a signal sequence preceding the recognition site for polyadenylation and/or polymerase II termination has been well established (40). However, the analysis of numerous cloned genes has brought new insights in the complexity of the factors involved in eukaryotic mRNA termination (41). In the 845 nucleotides of the 3′-untranslated region of the human pro-α2(I) collagen gene, we have found several potential signals for polyadenylation (Fig. 6), but only four of them are clearly utilized by the mature fibroblast transcripts. Moreover, two of these signals appear to be variations of the canonical sequence. The first, AUUAAA, has been reported to be a functional site for the mouse pancreatic α-amylase mRNA (42). The second, AUUAA, is a shorter version of this canonical variation. It is interesting to note that the 750-nucleotide dihydrofolate reductase mRNA also utilizes a shorter version (AUAA) of the canonical signal (AAUAAA) (16, 43). We have not yet tested if the expression and modulation of these transcripts varies under different conditions of cell culture or in tissues differentially expressing type I procollagen. In any event, it appears clear that the maturation of the pro-α2(I) collagen mRNA, besides the intrinsic complexity of the numerous splicing events, is further complicated by the generation of different size products. Our data strongly suggest that the differences are primarily due to the length of 3′-untranslated region. It could be argued that these multiple transcripts represent monogenic products of three or more pro-α2(I) collagen genes. This is a most unlikely explanation because biochemical, genetic, and molecular evidence has strongly suggested the presence of only one

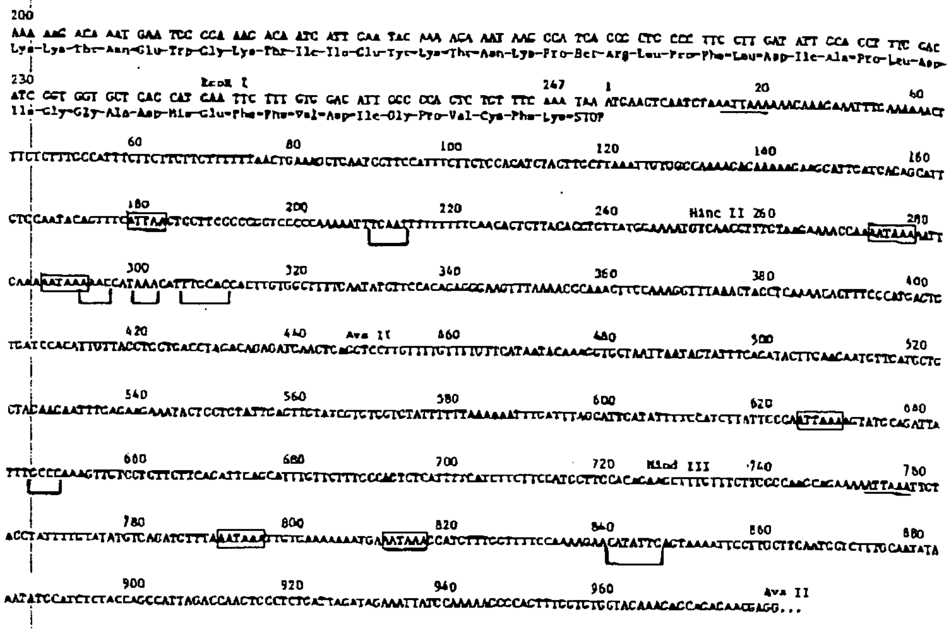

FIG. 6. **Nucleotide sequence of exon 1 of the human pro-α2(I) collagen gene and its immediate 3′-flanking region.** The last 48 amino acid codons of the COOH-terminal propeptide are numbered 200–247. The nucleotides of the 3′-untranslated region of exon 1 and the adjacent 3′-flanking regions are numbered 1–983. Some of the restriction enzyme sites are indicated in reference to the maps shown in Figs. 1, 3, and 5. Polyadenylation signals utilized by the fibroblast transcripts are *boxed*; other potential signals are *underlined*. The *bracket beneath* the nucleotide sequences indicates the approximate termination points of the four pro-α2(I) collagen mRNAs.

copy of the pro-α2(I) gene in the human haploid complement (34, 44, 45). Formally, it is also possible that although the pro-α2(I) gene is present in a single copy, its 3′ end may be shared by other genes, generating similar size transcripts. However, Southern blotting analysis of nuclear DNAs from different individuals digested with various restriction enzymes and hybridized to the *Eco*RI:*Bam*I genomic subclone clearly showed a unique pattern of single copy representation (data not shown).

Finally, it is tempting to speculate that, in some inherited or acquired disorders of connective tissue, an altered expression of one or more of these transcripts due to mutations at any of the control levels may result in a change in the production of functional pro-α2(I) mRNA.

Acknowledgments—We wish to thank Dr. T. Maniatis for his kind gift of the human genomic library, Dr. A. Bank for his helpful advice, Dr. D. J. Prockop for his enthusiastic support, and most of all Dr. D. Kaback for his invaluable help and kindness in the electron microscope studies.

REFERENCES

1. Bornstein, P., and Sage, H. (1980) *Annu. Rev. Biochem.* **49**, 957–1004
2. Prockop, D. J., Kivirikko, K. I., Tuderman, L., and Guzman, N. A. (1979) *N. Engl. J. Med.* **301**, 13–23; 77–85
3. Junien, C., Weil, D., Myers, J. C., Van Cong, N., Chu, M. L., Foubert, C., Gross, M. S., Prockop, D. J., Kaplan, J. C., and Ramirez, F. (1982) *Am. J. Hum. Genet.* **34**, 381–387
4. Huerre, C., Junien, C., Weil, D., Chu, M.-L., Morabito, M., Van Cong, N., Myers, J. C., Foubert, C., Gross, M.-S., Prockop, D. J., Boué, A., Kaplan, J. C., de la Chapelle, A., and Ramirez, F. (1982) *Proc. Natl. Acad. Sci. U. S. A.* **79**, 6627–6630
5. Myers, J. C., Chu, M.-L., Faro, S. H., Clark, W. J., Prockop, D. J., and Ramirez, F. (1981) *Proc. Natl. Acad. Sci. U. S. A.* **78**, 3516–3520
6. Chu, M.-L., Myers, J. C., Bernard, M. P., Ding, J.-F., and Ramirez, F. (1982) *Nucleic Acids Res.* **10**, 5925–5934
7. Bernard, M. P., Myers, J. C., Chu, M. L., Ramirez, F., Eikenberry, E. F., and Prockop, D. J. (1983) *Biochemistry* **22**, 1139–1145
8. Fuller, F., and Boedtker, H. (1981) *Biochemistry* **20**, 996–1006
9. Dickson, L. A., Ninomiya, Y., Bernard, M. P., Pesciotta, D. M., Parsons, J., Green, G., Eikenberry, E. F., de Crombrugghe, B., Vogeli, G., Pastan, I., Fietzek, P. P., and Olsen, B. R. (1981) *J. Biol. Chem.* **256**, 8407–8415
10. Wilson, A. C., Carlson, S. S., and White, T. J. (1977) *Annu. Rev. Biochem.* **46**, 573–639
11. Ohkubo, H., Vogeli, G., Mudryj, M., Avvedimento, V. E., Sullivan, M., Pastan, I., and de Crombrugghe, B. (1980) *Proc. Natl. Acad. Sci. U. S. A.* **77**, 7059–7063
12. Boyd, C. D., Tolstoshev, P., Schafer, M. P., Trapnell, B. C., Coon, H. C., Kretschmer, P. J., Nienhuis, A. W., and Crystal, R. G. (1980) *J. Biol. Chem.* **255**, 3212–3220
13. Wozney, J., Hanahan, D., Tate, V., Boedtker, H., and Doty, P. (1981) *Nature (Lond.)* **294**, 129–135
14. Monson, J. M., and McCarthy, B. J. (1981) *DNA* **1**, 59–69
15. Tosi, M., Young, R. A., Hagenbüchle, O., and Schibler, U. (1981) *Nucleic Acids Res.* **9**, 2313–2323
16. Setzer, D. R., McGrogan, M., and Schimke, R. T. (1982) *J. Biol. Chem.* **257**, 5143–5147
17. Unterman, R. D., Lynch, K. R., Nakhasi, H. L., Dolin, K. P., Hamilton, J. W., Cohn, D. V., and Feigelson, P. (1981) *Proc. Natl. Acad. Sci. U. S. A.* **78**, 3478–3482
18. Hellig, R., Perrin, F., Gannon, R., Mandel, J. L., and Chambon, P. (1980) *Cell* **20**, 625–637
19. Bennetzen, J. L., and Hall, B. D. (1982) *J. Biol. Chem.* **257**, 3018–3025
20. Young, R. A., Hagenbuchle, O., and Schibler, U. (1981) *Cell* **23**, 451–458
21. Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G., and Maniatis, T. (1978) *Cell* **15**, 1157–1174
22. Blattner, F. R., Williams, B. G., Blechl, A. E., Thompson, K. D., Faber, H. E., Furlong, L. A., Grunwald, D. J., Kiefer, D. O., Moore, D. D., Schumn, J. W., Sheldon, E. L., and Smithies, O. (1977) *Science (Wash. D. C.)* **196**, 161–169
23. Burnett, W., and Rosenbloom, J. (1979) *Biochem. Biophys. Res. Commun.* **86**, 478–484
24. Maxam, A. M., and Gilbert, W. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74**, 560–564
25. Thomas, P. S. (1980) *Proc. Natl. Acad. Sci. U. S. A.* **77**, 5201–5205
26. Southern, E. M. (1975) *J. Mol. Biol.* **98**, 503–517
27. Kaback, D. B., Angerer, L. M., and Davidson, N. (1979) *Nucleic Acids Res.* **6**, 2499–2517
28. Davis, R. W., Simon, M., and Davidson, N. (1971) *Methods Enzymol.* **21**, 413–428
29. Berk, A. J., and Sharp, P. A. (1978) *Proc. Natl. Acad. Sci. U. S. A.* **75**, 1274–1278
30. Tate, V., Finer, M., Boedtker, H., and Doty, P. (1982) *Cold Spring Harbor Symp. Quant. Biol.*, in press
31. Gilbert, W. (1978) *Nature (Lond.)* **271**, 501
32. Monson, J. M., Friedman, J., and McCarthy, B. J. (1982) *Mol. Cell. Biol.* **2**, 1362–1371
33. Barsh, G. S., and Byers, P. H. (1981) *Proc. Natl. Acad. Sci. U. S. A.* **78**, 5142–5146
34. Byers, P. H., Siegel, R. C., Peterson, K. E., Rowe, D. W., Holbrook, K. A., Smith, L. T., Chang, Y. H., and Fu, J. C. C. (1981) *Proc. Natl. Acad. Sci. U. S. A.* **78**, 7745–7749
35. de Wet, W. J., Pihlajaniemi, R., Myers, J. C., Kelly, T. E., and

Prockop, D. J. (1983) *J. Biol. Chem.* **258**, 7721-7728
36. Schmid, C. W., and Jelinek, W. R. (1982) *Science (Wash. D. C.)* **216**, 1065-1070
37. Shapiro, J. A. (1979) *Proc. Natl. Acad. Sci. U. S. A.* **76**, 1933-1937
38. Rogers, J., Early, P., Carter, C., Calame, K., Bond, M., Hood, L., and Wall, R. (1980) *Cell* **20**, 303-312
39. Amara, S. G., Jonas, V., Rosenfeld, M. G., Ong, E. S., and Evans, R. M. (1982) *Nature (Lond.)* **298**, 240-244
40. Proudfoot, N. J., and Brownlee, G. G. (1976) *Nature (Lond.)* **263**, 211-214
41. Proudfoot, N. J. (1982) *Nature (Lond.)* **298**, 516-517
42. Hagenbuchle, O., Bovey, R., and Young, R. A. (1980) *Cell* **21**, 179-187
43. Setzer, D. R., McGrogan, M., Nunberg, J. H., and Schimke, R. T. (1980) *Cell* **22**, 361-370
44. Steinmann, B., Tuderman, L., Peltonen, L., Martin, G. R., McKusick, V. A., and Prockop, D. J. (1980) *J. Biol. Chem.* **255**, 8887-8893
45. Dalgleish, R., Trapnell, B. C., Crystal, R. G., and Tolstoshev, P. (1982) *J. Biol. Chem.* **257**, 13816-13822